



Cornell University  
ILR School

Cornell University ILR School  
**DigitalCommons@ILR**

---

Labor Dynamics Institute

Centers, Institutes, Programs

---

5-3-2013

## Presentation: Improved Research Access to Census Bureau Linked Administrative Data via Public-use Products

John Abowd

*Cornell University*, [John.Abowd@cornell.edu](mailto:John.Abowd@cornell.edu)

Lars Vilhuber

*Cornell University*, [lv39@cornell.edu](mailto:lv39@cornell.edu)

Follow this and additional works at: <https://digitalcommons.ilr.cornell.edu/ldi>

Thank you for downloading an article from DigitalCommons@ILR.

**Support this valuable resource today!**

---

This Article is brought to you for free and open access by the Centers, Institutes, Programs at DigitalCommons@ILR. It has been accepted for inclusion in Labor Dynamics Institute by an authorized administrator of DigitalCommons@ILR. For more information, please contact [catherwood-dig@cornell.edu](mailto:catherwood-dig@cornell.edu).

If you have a disability and are having trouble accessing information on this website or need materials in an alternate format, contact [web-accessibility@cornell.edu](mailto:web-accessibility@cornell.edu) for assistance.

---

## **Presentation: Improved Research Access to Census Bureau Linked Administrative Data via Public-use Products**

### **Abstract**

Professor John M. Abowd, Director of the Labor Dynamics Institute, presented on "Improved Research Access to Census Bureau Linked Administrative Data via Public-use Products" at the 18th Annual Meetings of the Society of Labor Economists in Boston, MA, on Friday, May 3, 2013.

# Improved Research Access to Census Bureau Linked Administrative Data via Public-use Products

John M. Abowd and Lars Vilhuber  
Cornell University

with a big assist from Abigail Cooke, Javier  
Miranda, Martha Stinson, and Kelly Trageser

May 3, 2013

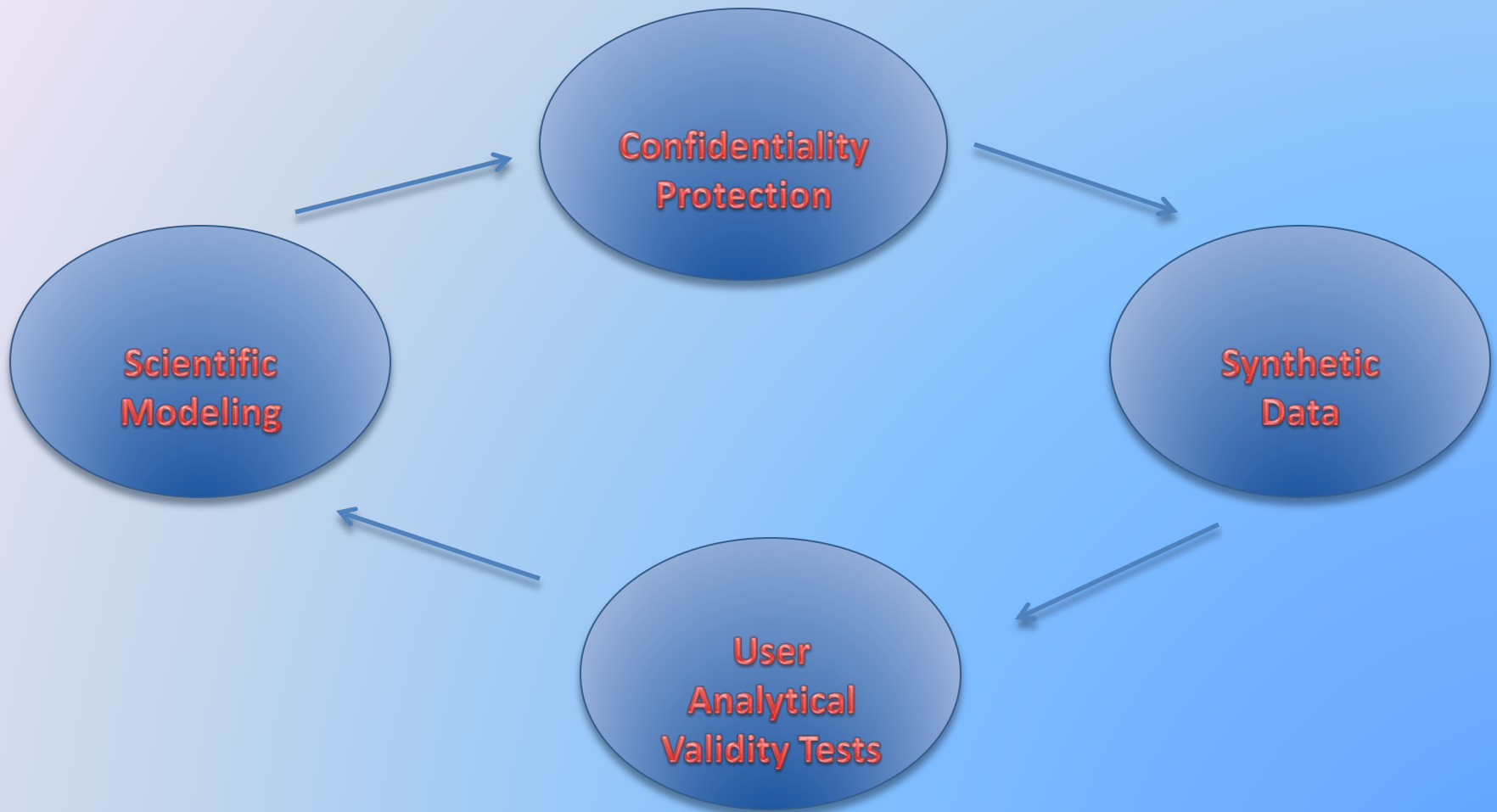
# Outline

- Why you should consider using synthetic data-  
*seriously*
- SIPP synthetic data
- LBD synthetic data
- OnTheMap/LODES data
- The Quarterly Workforce Indicators (not synthetic, but used a related technique)

# Why Use Synthetic Data

- Most simple analyses give answers that are useful
- Most complex designs do not
- But, the Census Bureau developers have a feedback loop that allows you to get results from the underlying confidential data
- Whole process takes weeks, not years
- New releases of the synthetic data reflect the features discovered by the users

# Feedback Loop



# **SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP) SYNTHETIC DATA**

# Genesis of the Synthetic SIPP

- Developed originally by a portion of the SIPP user community primarily interested in national retirement and disability programs
- SIPP augmented with
  - earnings histories from the IRS data maintained at SSA (W-2)
  - benefit data from SSA's master beneficiary records.
- Feasibility assessment (confidentiality!) of adding SIPP variables to earnings/benefit data in a public-use file (PUF)
  - set of variables that could be added without compromising the confidentiality protection of the existing SIPP public use files was VERY limited
- Alternative methods explored



# Basic Methodology

- Experiment using synthetic data (Rubin, 1993; Little, 1993)
- Partially synthetic data with multiple imputation of missing items (Reiter, 2004)
- Partially synthetic data:
  - Some variables are actual responses
  - Other variables are replaced by values sampled from the posterior predictive distribution for that record, conditional on all of the confidential data

# Brief History of the Synthetic SIPP

- 2003-2005: Creation, but no release, of three versions of the “SIPP/SSA/IRS-PUF” (SSB)
- 2006: [Release to limited public access of SSB V4.2](#)
  - Access to general public only at Cornell-hosted Virtual RDC (SSB server: restricted-access setup)
  - Promise of evaluation of Virtual RDC-run programs on internal Gold Standard
  - Ongoing SSA evaluation
  - Ongoing evaluation at Census Bureau (including RDCs)
- 2010: Release of [SSB V5](#) at Census and on the VirtualRDC (Synthetic Data Server, [SDS](#))
  - Codebook: [http://www.census.gov/sipp/ssb/SSB\\_Codebook.pdf](http://www.census.gov/sipp/ssb/SSB_Codebook.pdf)
  - Restructured to vastly improve analytical validity of SIPP variables
- 2013: Release of SSB V5.1 at Census and on the Cornell VirtualRDC (SDS)
  - Documentation in preparation (meta-data in CED<sup>2</sup>AR)
  - First user-initiated variables

# Basic Structure of the SSB V4

- SIPP
  - Core set of 125 SIPP variables in a standardized extract of SIPP panels 1990-1993 and 1996
  - All missing data items (except for structurally missing) are marked for imputation
- IRS
  - Maintained at SSA, but derived from IRS records
  - Master summary earnings records (SER)
  - Master detailed earnings records (DER)

# Basic Structure of the SSB V4 (II)

- SSA
    - Master Beneficiary Record (MBR)
  - Census
    - Numident: administrative birth and death dates
  - All files combined using verified SSNs
- => “Gold Standard”

# Basic Structure of SSB V5

- Panels: 1990, 1991, 1992, 1993, 1996, 2001, and 2004 (this variable is now in the SSB)
- Couple-level linkage: the first person to whom the SIPP respondent was married during the time period covered by the SIPP panel
- SIPP variables only appear in years appropriate for the panel indicated by the PANEL variable (biggest change from V4.2)

# SIPP Synthetic Beta V5.1

- Available (really soon) through the Cornell Virtual RDC
- Includes 7 SIPP panels
  - 1990, 1991, 1992, 1993, 1996, 2001, 2004
- Administrative earnings history was edited to remove some erroneous reports prior to synthesis
- Adds many new SIPP variables that vary across months of the panel years: (user-requested variables: feedback loop!)
  - Weeks with job, weeks with pay
  - Usual hours worked
  - Survey-reported earnings
  - Total personal income
  - Health insurance coverage, any
  - Health insurance coverage, employer-provided
- Includes state of residence recorded in first SIPP interview

# SIPP Synthetic Beta V6.0

- Expected Release: May 2014
- Added 1984 and 2008 panels
- Extend administrative earnings history through 2011
- Include children (under age 15)
- Include more SIPP and administrative disability variables
- Include a few characteristics of employers



# Ongoing Research

- Developing job-level Gold Standard File as companion to person-level file. Contains earnings by employer from 1978-2011.
- Developing link between parents and children in the Gold Standard File. (See also [Gottschalk and Stinson](#), SOLE 2013 Friday 8AM, Beacon H)
- Research on how to synthesize these links to protect confidentiality so these data could be released.



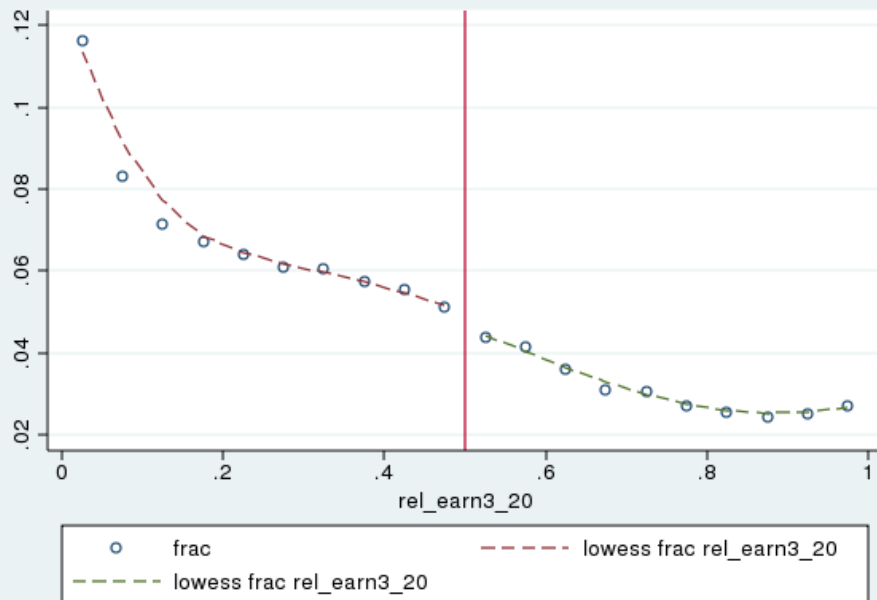
# External Researcher Validation

- Version 4
  - 12 projects
  - 1 was submitted for validation
- Version 5
  - 31 projects
  - 6 were submitted for validation

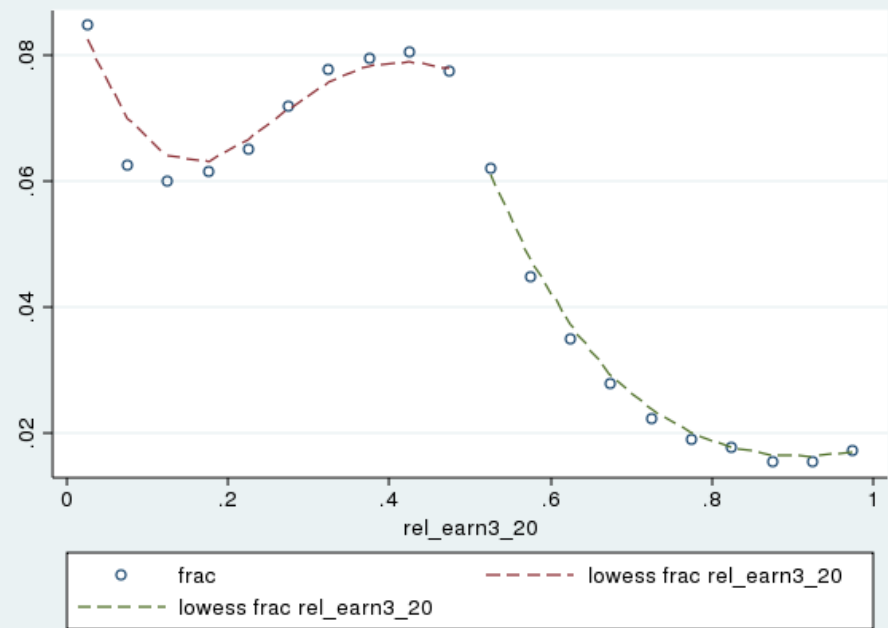
# Validation Details

- Henriques, Alice (2102) “How does Social Security claiming respond to incentives? Considering husbands’ and wives’ benefits separately”
- Armour, Philip (2012) “The role of information in disability insurance take-up: An analysis of the Social Security statement phase-in”
- Bertrand, Marianne, Emir Kamenica and Jessica Pan (2013) “Gender identity and relative income within households”

# From Bertrand *et al.*



data source: ssb\_v5\_0\_synthetic:1\_1.dta



Timeline: SDS application November 2012, gold standard results January 2013

# CED<sup>2</sup>AR

About CED<sup>2</sup>AR | Login or Register

## CED<sup>2</sup>AR

The Comprehensive Extensible Data Documentation and Access Repository

### Welcome to CED<sup>2</sup>AR!

You currently have no search filters set. You may limit your searches by a given codebook below:

- [SSB](#)
- [IPUMS USA](#)

### Simple Search

### Advanced Search

### Browse Metadata

Enter keywords below to do a broad search of ALL FIELDS within the available codebook metadata.  
(Hint: For a more refined search, use the [Advanced Search](#) form.)

Search

### Welcome to the Comprehensive Extensible Data Documentation and Access Repository (CED<sup>2</sup>AR)

CED<sup>2</sup>AR is a [National Science Foundation \(NSF\) funded](#) project developed by the [NSF Census Research Network - Cornell Node \(NCRN\)](#).

It is designed to improve the documentation and discoverability of both public and restricted data from the federal statistical system.

To search across all datasets in the repository, enter a term in the search box above and click the 'Search' button.

# CED<sup>2</sup>AR

The screenshot displays the CED<sup>2</sup>AR website interface. At the top, the header includes the CED<sup>2</sup>AR logo and the tagline "The Comprehensive Extensible Data Documentation and Access Repository", along with links for "About CED<sup>2</sup>AR" and "Login or Register". A modal window titled "Document Description" is open, showing details for the "SSB" dataset. The modal includes sections for "Citation" (Title Statement, Title: SSB, Production Statement) and a note that "95 variables found." The background shows a sidebar with a message about the current dataset (SSB) and a main content area with a search box and a description of CED<sup>2</sup>AR as an NSF-funded project.

**CED<sup>2</sup>AR** The Comprehensive Extensible Data Documentation and Access Repository

About CED<sup>2</sup>AR | Login or Register

You are currently viewing metadata for [SSB](#).

To view other datasets' metadata, [remove this filter](#).

**Document Description**

**Citation**

Title Statement  
Title: SSB  
Production Statement

95 variables found.

CED<sup>2</sup>AR is a [National Science Foundation \(NSF\) funded](#) project developed by the [NSF Census Research Network - Cornell Node \(NCRN\)](#).

It is designed to improve the documentation and discoverability of both public and restricted data from the federal statistical system.

To search across all datasets in the repository, enter a term in the search box above and click the 'Search' button.

May 3, 2013

Abowd and Vilhuber

# CED<sup>2</sup>AR

About CED<sup>2</sup>AR | Login or Register

**CED<sup>2</sup>AR** The Comprehensive Extensible Data Documentation and Access Repository

You are currently viewing metadata for [SSB](#).

To view other datasets' metadata, [remove this filter](#).

Simple Search

Advanced Search

Browse Metadata

## Browse Alphabetically

[A](#) | [B](#) | [C](#) | [D](#) | [E](#) | [G](#) | [H](#) | [I](#) | [J](#) | [K](#) | [L](#) | [M](#) | [N](#) | [O](#) | [P](#) | [Q](#) | [R](#) | [S](#) | [T](#) | [U](#) | [V](#) | [W](#) | [X](#) | [Y](#) | [Z](#)

4 results returned.

[<<Search again](#)

Show  entries

Search:

Variable ▲	Label ▼
<a href="#">hicov</a>	Health Insurance Coverage
<a href="#">hiemp</a>	Health Insurance Coverage from Employer
<a href="#">hispanic</a>	Hispanic
<a href="#">homeequity</a>	Home Equity

Showing 1 to 4 of 4 entries

◀ [Previous](#) [Next](#) ▶

# CED<sup>2</sup>AR

**CED<sup>2</sup>AR** The Comprehensive Extensible Data Documentation and Access Repository

[About CED<sup>2</sup>AR](#) | [Login or Register](#)

You are currently viewing metadata for [SSB](#).

To view other datasets' metadata, [remove this filter](#).

Simple Search

Advanced Search

Browse Metadata

[<< Back To List](#)

## hicov

<b>Codebook:</b>	SSB
<b>Producer:</b>	
<b>Variable Name:</b>	hicov
<b>Label:</b>	Health Insurance Coverage
<b>Var Type:</b>	numeric
<b>Values:</b>	0 Respondent did not have health insurance coverage during this month 1 Respondent had health insurance coverage during this month Sysmiss

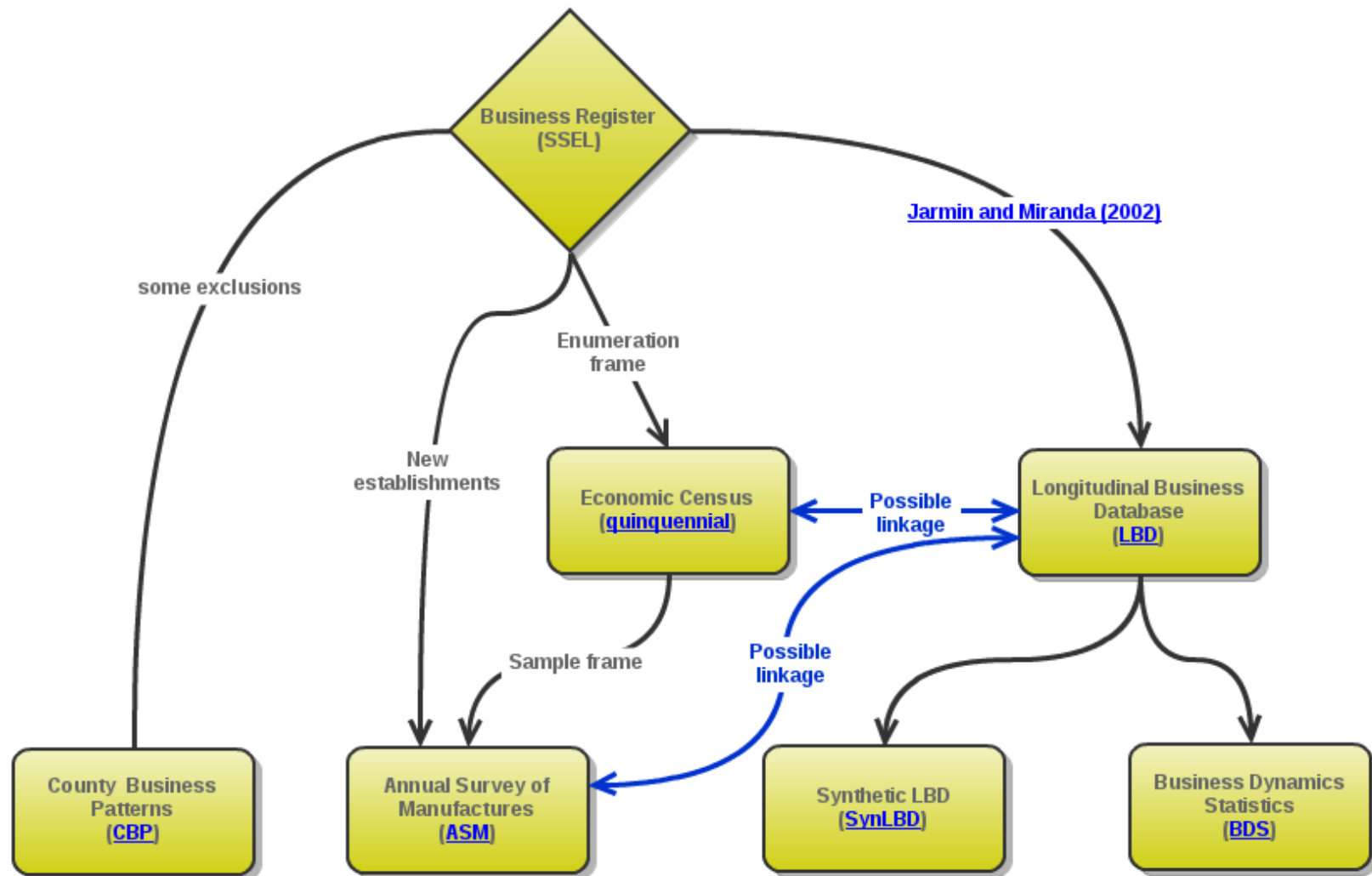
# CED<sup>2</sup>AR

- Part of the NSF-Census Research Network
- Funded through NSF Grant [#1131848](#)
- Soon live, watch <http://www.ncrn.cornell.edu> (Cornell node) or <http://www.ncrn.info> (Network)



# **SYNTHETIC LONGITUDINAL BUSINESS DATABASE**

# LBD Provenance



# Longitudinal Business Database

- Economic census covering nearly all private non-farm business establishments with paid employees
  - Annual payroll and Mar 12 employment (1976-2011)
  - SIC/NAICS
  - Geography (down to county),
  - Entry year, Exit year
  - Firm structure
- Used for looking at business dynamics, job flows, market volatility, international comparisons
- Most widely-used Census Bureau data set in RDCs

# Synthetic LBD

- Synthetic data set
  - [Available outside the Census RDC on Cornell VirtualRDC Synthetic Data Server](#)
  - Provides analytical validity for dynamic establishment-level employment and payroll
  - Reduces the number of requests for special tabulations from the LND
  - Aids users applying for RDC access
- Experiment in public-use business micro-data

# Creation of the Synthetic LBD

- Version 2.0 described in [Kinney et al \(2012\)](#) and <https://www.census.gov/ces/dataproducts/synlbd/index.html>
- Partially synthetic data
  - Employment, payroll, birth and death date of establishments, multi-unit status
  - Suppressed county, SIC released
  - 1976-2000

# External Validation Exercises

- 41 approved projects (includes provisional approvals)
- 3 have submitted results for validation (one of these did two rounds of validation)
- One timeline: application approved March 2011, validation results released September 2011
- 14 projects denied because data not available (typically wanted NAICS, geography, firm id)

# Validation Exercise Details

- Professor: fixed-effects regressions of establishment-level employment growth interacted with unemployment rate, by age, sector, and single/multi (into 2<sup>nd</sup> round)
- Ph.D. student: turnover rates of establishments by industry, model with implications on concentration and exit/entry dynamics
- Ph.D. student: shape of the distribution of entering establishments at different points in time (into 3<sup>rd</sup> round)

# Ongoing Development Efforts

- SynLBD 2.x ( $x > 0$ )
  - Same methodology as 2.0
  - Adds additional years (through 2011)
  - Based on NAICS (using back-coding by Shawn Klimek and [Teresa Fort](#))
  - Data created, working on disclosure avoidance documentation
- SynLBD 3.0
  - Ongoing work by Kinney, Reiter
  - Adds synthesis of firm links, geography, based on NAICS



# Access to SynLBD

- See

<https://www.census.gov/ces/dataproducts/synlbd/index.html> and

<http://www.vrdc.cornell.edu/sds/>

# **OTHER CENSUS BUREAU SYNTHETIC DATA PRODUCTS**

# OnTheMap/LODES

- [OnTheMap](#)
  - Workplace and residence coded to census block
  - Data for 2002-2010 (2011 soon)
  - Residence address is synthetic
  - All queries explicitly logged (permitting analysis on underlying confidential data by Census staff)
- [LODES](#)
  - Public-use data in tabular format

# Quarterly Workforce Indicators

- Local labor market time series on employment, earnings, hiring, separation, job creations, job destructions
- Simpler confidentiality protection system (noise infusion)
- Based on linked unemployment insurance and QCEW data
- Age, sex, race, ethnicity, education, ownership, NAICS, CBSA, firm size, firm age all available for all variables

# Lots More Information

- [Cornell University Information Science 7470 Understanding Social and Economic Data](#)
- Partially sponsored by the NSF-Census Research Network
- All materials, exercises, lecture videos online